**DOCUMENT RESUME**

ED 161 930 TM 007 957

AUTHOR Plake, Barbara S.; Hoover, H. D.
TITLE A Methodology for Identifying Biased Achievement Test Items that Removes the Confounding in an Items by Groups Interaction Due to Possible Group Differences in Instructional Level.
PUB DATE [Mar 78]
NOTE 7p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Achievement Tests; *Analysis of Variance; *Complexity Level; *Item Analysis; Sampling; *Test Bias; Test Items

ABSTRACT

A method of investigating for possible bias in test items is proposed that uses analysis of variance for item data based on groups that have been selected to have identical test score distributions. The item data used are arcsin transformations of item difficulties. The methodological procedure has the following advantages: (1) The arcsin transformation stabilizes the variance within replications. (2) Since the groups are formed to have equal distributions of raw scores, there is no achievement difference between the groups. (3) Since the selection procedure allows for consideration of the entire raw score distribution, a range of raw scores is represented within each group. (4) The use of variance components allows for an investigation of the practical significance of the variance components; and (5) The follow-up procedure is in the same units as the analysis of variance, provides Type I error control, and investigates the items by contributing to a significant items-by-group interaction. (Author/CTM)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED161930

# A Methodology for Identifying Biased Achievement Test Items that Removes the Confounding in an Items by Groups Interaction Due to Possible Group Differences in Instructional Level

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Barbara S. Plake
University of Nebraska-Lincoln

H. D. Hoover
University of Iowa

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Barbara S. Plake
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

An item in a test may be considered biased if it functions differentially for groups of students. One technique of screening a test for biased items is an analysis of variance procedure. In studies using ANOVA procedures, the significant items by groups interaction is generally used as a signal that biased items may be present. Follow-up procedures are typically used to identify the item(s) contributing to the items by group interaction.

A variety of analysis of variance models and techniques have been suggested for screening a test for biased items (Cardall and Coffman, 1964, Cleary and Hilton, 1968, Angoff and Sharon, 1972). The methodology reported in these studies has the weakness that a significant items by groups effect may be due to differing instructional levels between the two groups. This may confound the presence and interpretation of an item by groups interaction.

TM007 957

One of the reasons for checking an achievement test for biased items is to remove items that are systematically favoring one group of examinees differently than would be expected by initial achievement level. Examination of achievement test items for race and sex bias is currently contemplated by major achievement test authors. Race bias investigation in achievement tests has the potential confounding problem identified as a weakness in the general ANOVA model for screening a test for bias properties. Within any particular grade level, the instructional level of black and white examinees may differ substantially, especially in big cities. Therefore, to form the two race groups for a grade level has the definite possibility of creating two groups that differ substantially in instructional experiences and level.

Generally, an analysis of variance model with q groups (race, sex) and r replications (school systems, cities, states) used in the study of a test with p items has the following form:

$$\theta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijk}$$

where:

$\theta_{ijk}$ = arcsin transformed item difficulty for the $i^{th}$ item, $j^{th}$ group and $k^{th}$ replication.

$\mu$ = overall mean

$\alpha_i$ = item effect ($i = 1,2,\ldots,p$)

$\beta_j$ = group effect ($j = 1,2,\ldots,q$)

$\gamma_k$ = replication effect ($k = 1,2,\ldots,r$)

$(\alpha\beta)_{ij}$ = the interaction of the $i^{th}$ item with the $j^{th}$ group

$(\alpha\gamma)_{ik}$ = the interaction of the $i^{th}$ item with the $k^{th}$ replication.

$(\beta\gamma)_{jk}$ = the interaction of the $j^{th}$ group and the $k^{th}$ replication.

$(\alpha\beta\gamma)_{ijk}$ = the interaction of the $i^{th}$ item with the $j^{th}$ group and the $k^{th}$ replication.

By using a selection rule to systematically select group members, it is possible to negate the between group instructional level difference. One possible selection rule would be to select members for the groups systematically so that each group has members that have identical raw score distributions. It is attractive to sample group members for both groups from the entire distribution of raw scores (if possible). Then the groups will have a wide distribution of raw scores within but will have identical distributions of raw scores between. Notice this selection procedure allows for the generation of groups that are equal in size and have identical distributions of raw scores. The analysis of variance model then reduces to:

$$\theta_{ij} = \mu + \alpha_i + \gamma_k + (\alpha\gamma)_{ik} + (\alpha\beta)_{ij} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijk}$$

since the $\beta$; and $(\beta\gamma)_{jk}$ terms are zero due to the matching of the groups on total score.

The hypothesis of interest is: $H_o$: $\sigma^2_{\alpha\beta} = 0$, that is, there is no interaction between items and groups. With a large sample size and/or items in a test, the large degrees of freedom will

allow for increased power. Therefore, in order to ascertain the practical potence of the items by groups component, the F ratio should be coupled with a consideration of the variance components.

When a significant items by groups interaction is found, a follow-up procedure should be used to identify the items contributing to the items by group interaction. Plake and Hoover (1977) recommend the use of a Bonferroni follow-up procedure that is based on the interaction contrast. This follow-up procedure can be expressed in arcsin transformed item difficulty units and is, therefore, in the same units as the ANOVA. The Bonferroni follow-up provides adequate experiment wise Type I error control without sacrificing desirable amounts of power. Finally, the follow-up procedure directly investigates the interactive contrast for biased items. One contrast is obtained for each item in the test. This series of contrasts can be presented as follows:

$$\Psi_{ij} = \theta_{ij} - \bar{\theta}_{i} - \bar{\theta}_{j} + \bar{\bar{\theta}}$$

where:

$\Psi_{ij}$ = interaction contrast value for the $i^{th}$ item and $j^{th}$ group across the r replications

$\theta_{ij}$ = arcsin transformed item difficulty for the $i^{th}$ item and $j^{th}$ group averages across the r replication

$\bar{\theta}_{i}$ = mean arcsin transformed item difficulty across the q groups and the r replications

$\bar{\theta}_{j}$ = mean arcsin transformed item difficulty across the p items and r replications

$\bar{\bar{\theta}}$ = overall mean

Since these contrasts can be expressed as a linear combination of cell means, under the assumptions of the ANOVA, they are distributed normally with variance estimated by:

$$\sigma^2_{\Psi_{ij}} = \frac{MS_{error}}{r}\left[\sum_{M=1}^{p+q+a} C_M^2\right]$$

where $MS_{error}$ is the appropriate error term for the items by groups interaction component and $C_M$ is the coefficient of the $M^{th}$ term of the linear expression of cell means that defines the particular interaction contrast.

The methodological procedure outlined above has the following advantages:

1) The arcsin transformation stabilizes the variance within replications.

2) Since the groups are formed to have equal distributions of raw scores, there is no achievement difference between the groups.

3) Since the selection procedure allows for consideration of the entire raw score distribution, a range of raw scores is represented within each group.

4) The use of variance components allows for an investigation of the practical significance of the variance components.

5) The follow-up procedure is in the same units as the ANOVA, provides Type I error control, and investigates the items contributing to a significant items by group interaction.

## Bibliography


Angoff, W. H. and Sharon, S. F., Item-race interaction on a test of scholastic aptitude, RB-71-59, Educational Testing Service, Princeton, New Jersey, 1972.

Cardell, C., and Coffman, W. E., A method of comparing the performance of different groups on the items in a test, RB-64-61, Educational Testing Service, Princeton, New Jersey, 1964.

Cleary, T. A., and Hilton, T. L., An investigation of item bias, Educational and Psychological Measurement, 1968, 1, 61-75.

Plake, B. S., and Hoover, H. D., An analytical method of identifying biased test items. Paper presented at the American Educational Research Association annual meeting, April, 1977.